DOCUMENT RESUME

ED 271 507 TM 860 418

AUTHOR Koon, Jeff
TITLE The Underevaluation of Student Ratings of Academic Programs.
PUB DATE Apr 86
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
AVAILABLE FROM Jeff Koon, 1662 Blair Avenue, St. Paul, MN 55104 ($2.00 ea., $15.00 for 10).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Correlation; Higher Education; *Interrater Reliability; Program Effectiveness; *Program Evaluation; Questionnaires; *Rating Scales; Student Attitudes; *Student Evaluation of Teacher Performance; *Surveys
IDENTIFIERS F Test; *Mail Surveys; Pearson Product Moment Correlation

ABSTRACT

This study examines the general hypothesis that surveys can be used to elicit reliable and valid student ratings of the effectiveness of most aspects of the delivery of the academic programs in which they are enrolled, and reports data on test-retest reliability for upper division majors as well as on the reliability with which ratings differentiated between programs at one institution. Strict behavioral measures have important limitations in academic program evaluation while survey items can be fitted to almost any content. Related types of student ratings are reasonably valid. Fity-five majors in engineering and the physical sciences had a mean reliability of .66 (Pearson r) when re-rating 58 aspects of their 12 academic programs over a three-month interval. The mean test-retest reliability for 58 items was .57, and the item means were stable over time (F-test). Mean ratings by students (N=620) in the 12 programs, and by students (N=496) in 14 social science-related programs, often discriminated effectively between two or more programs. The results support the hypothesis, particularly since aggregation of survey ratings rapidly increases their reliability. The need for student ratings in reviews of educational quality, and for further study of potential biases in them, is outlined. A three-page list of references concludes the report. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED271507

THE UNDEREVALUATION OF STUDENT RATINGS OF ACADEMIC PROGRAMS

Jeff Koon
University of California, Berkeley

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Koon

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A paper presented at the Annual Meeting of the
American Educational Research Association.
San Francisco, April 1986.

TM 860 414

Abstract

THE UNDEREVALUATION OF STUDENT RATINGS OF ACADEMIC PROGRAMS

Jeff Koon

The author hypothesizes that surveys can be used to elicit reliable and valid student ratings of the effectiveness of most aspects of the delivery of the academic programs in which they are enrolled, and reports data on test-retest reliability for upper division majors as well as on the reliability with which ratings differentiated between programs at one institution. Strict behavioral measures have important limitations in academic program evaluation while survey items can be fitted to almost any content. Related types of student ratings are reasonably valid.

Fifty-five majors in engineering and the physical sciences had a mean reliability of .66 (Pearson r) when re-rating 58 aspects of their 12 academic programs over a three-month interval. The mean test-retest reliability for 58 items was .57, and the item means were stable over time (F-test). Mean ratings by students (N=620) in the 12 programs, and by students (N=496) in 14 social science-related programs, often discriminated effectively between two or more programs. The results support the hypothesis, particularly since aggregation of survey ratings rapidly increases their reliability. The need for student ratings in reviews of educational quality, and for further study of potential biases in them, is outlined.

THE UNDEREVALUATION OF STUDENT RATINGS OF ACADEMIC PROGRAMS

Jeff Koon
1662 Blair Ave.
St. Paul, MN 55104

The potential for student ratings to contribute to the evaluation of academic programs is the subject of this paper. In evaluating academic programs, three related types of evaluations are potentially of special importance: student ratings of instructors and courses; student ratings of the effectiveness of the various aspects of program delivery; and student ratings of the extent of their own learning and development. In addition, the last two types may be combined in some important ways (see below).

Objectives

While the discussion will make reference to all three types of student evaluations, the general hypothesis, which pertains mainly to one, is that (appropriately designed) surveys can be used to elicit reliable and valid student ratings of the effectiveness of most aspects of the delivery of the academic programs in which they are enrolled. The objectives here are to portray the plausibility of this hypothesis and its implications for program evaluation theory, methods, and practice in higher education, and to report on a few of the measurement properties of upper division majors' ratings of various aspects of program delivery at one institution.

Certainly there are limits to the general hypothesis that student evaluations of their academic programs can be designed to be reliable and valid. The extent of students' knowledge of their subject matter, and of academic programming, for example, limit the questions that reasonably may be asked. The "political" climate (e.g., if retrenchment is expected) may affect the validity of student ratings. All types of threats to validity (c.f., Cook and Campbell, 1979) must be considered in estimating the typical extent and sources of bias. But since such ratings are to be aggregated, such that random error will wash out, it is not necessary that each student's ratings be highly reliable or valid (c.f., The Joint Committee on Standards for Educational Evaluation, 1981). As long as student ratings are somewhat reliable and valid, only non-random, extraneous biases that cannot be controlled or counterbalanced, and that are of a consequential magnitude, are a real threat.

Perspectives on the General Hypothesis

Currently, little is known about what works--about cause and effect relationships--in higher education (c.f., Astin, 1980). Yet improvements in academic programming, and effective evaluations, are contingent upon knowing which aspects of program delivery contribute to which outcomes, and the extent of their effects (c.f., Ball, 1981; Ewell, 1984).

In part, this lack of knowledge is due to current evaluative theory, methods, and practice, especially as applied to higher education. Since researchers can only very rarely use random selection or impose laboratory conditions in evaluating higher education, we shall assume in the discussion

to follow that students self-select their programs and that evaluative models that insist on experimental methods are almost useless.

Problems with Strict Behavioral Measures Perhaps the greatest problem in the traditional approach to evaluation is that strict behavioral measures often are deemed essential. The strict behavioral approach produces inadequate evaluations in part because it seldom is possible, let alone feasible, to select measurable objectives that provide reasonably comprehensive coverage of the effects on students of participating in an academic program. As a result, behaviorists often have let readily available measures "drive" their evaluations. Measures of non-cognitive aspects of student learning and development, and even of complex cognitive functioning, are seldom included in assessments, and there are only a few, often indirect measures for most aspects of program delivery.

The behavioral approach also is inadequate because the complexity of human social reality utterly overwhelms the attempt at concrete precision. Measures must be accurately fitted to each criterion of concern in an evaluation. An "indicator" or measure pertaining to some aspect of a criterion of concern cannot be accurate unless that aspect is an unbiased representative of the whole. Analysts must clearly identify a comprehensive set of outcomes, both expected and unexpected, and the various aspects of program delivery that contribute to them. Further, any student predisposition or activity (see below) that has a meaningful affect on an outcome, or changes the effects of an aspect of program delivery, needs to be controlled. Although a number of authors (c.f., Bowen, 1974; Clark, 1979; Solmon, 1981; Astin, 1985) warn that an evaluation is meaningless if it measures "outcomes" (really output levels) without controlling for inputs, only a few (e.g., Cronbach et al., 1980) warn also that the statistical methods for controlling those inputs are far from perfect. The analyst needs to estimate the direct and interactive effects on outcome from each aspect of program delivery--a task that is nearly impossible with self-selected groups in a field setting (Byrk, 1978).

Although the amount of influence on the student attributable to the academic department is important (Feldman and Newcomb, 1969; Hartnett and Centra, 1978), Everett K. Wilson estimates that 70% of what students learn while in college comes from other sources (Kuh, 1981, pp. 15, 19). Among the more important non-programmatic sources of influence on students are: the extracurriculum; the residential situation and the college community; peers; family; work and related responsibilities (e.g., leadership roles); the student's predispositions (values, attitudes, abilities, etc.), own intellectual interests and leisure time activities (c.f., Koon, 1974). In addition, learning attributable to coursework offered in other departments may need to be discounted. The direct and interactive effects of such influences must be controlled before the analyst can begin to surmise that the remaining variance in the output measurements are attributable to an academic program, let alone to particular aspects of its delivery.

The Potential of Student Ratings The general hypothesis is attractive in part because student evaluations do not share the host of weaknesses present and partly inherent in strict behavioral methods and measures (see also the comments on achievement tests, below). This alone should have led many academic program evaluators to study whether student

ratings are valid sources of information which currently is inadequate or unavailable. Indeed, if those most directly experienced can supply us with valid perceptions about most aspects of academic program delivery and about their own learning, student ratings could prove to be better than many of the measures currently in use.

There also is a crucial methodological consideration to take into account: survey items can be fitted almost exactly to any desired content. Students can be asked to rate the overall effectiveness of an academic program, or any aspect of program delivery. They can be asked to rate the extent of the contribution made by any aspect of program delivery to any outcome, while at the same time excluding the effects of other aspects of their program--thereby obviating the need for statistical controls. Similarly, they can be asked to distinguish between learning attributable to the academic program and to other aspects of their environment and selves. For indirect analyses, student ratings of their values, attitudes, interests, and abilities can assist in the identification of input characteristics that affect outputs and outcomes, and be used to attempt to understand how particular styles of academic program delivery affect different "types" of students. Such methodological properties (or potentials in survey design) are especially useful in a value-added analysis--and value-added necessarily is the concern when we seek to measure program effectiveness.

In sum, if the general hypothesis is substantially correct, student evaluations of academic programs could well be the missing link in evaluative theory and practice in higher education.

Evidence Pertaining to Related Uses of Student Evaluations Since student ratings of teaching are related closely to student ratings of the aspects of program delivery, the more extensive research on the former is germane here and will be reviewed briefly. Koon (1986) writes that a vast majority of researchers find that mean student ratings of instructor effectiveness are reasonably reliable and valid, and that potential biases (e.g., course/student level, the student's expected grade, class size) usually are of minor importance or can be subjected to statistical controls (c.f., Cohen, 1981; Marsh, 1980; McKeachie, 1979; Centra, 1979; Costin, Greenough and Menges, 1971).

Moreover, the correlation between student achievement and student ratings of teacher effectiveness, which Cohen (1981, p. 301) estimated at .43 in a meta-analysis, undoubtedly greatly understates the validity of the ratings (Koon, 1986). Previous studies of the validity of student ratings have relied almost exclusively on objective measures of learning as the criterion (c.f., Dowell and Neal, 1982; Benton, 1982). Measures of learning are a subset of a much greater whole; they rarely subsume the professor's contributions to student interest in a course or field, to the motivation to learn, to a number of abilities involved in analyzing and synthesizing information, to personal growth, to understanding of life, to the appreciation of civic or social responsibilities, to student retention of the material learned, and so on (Koon, 1986).

While there are many prospective criterion measures that have not been studied, Benton (1982), in summarizing findings pertaining to the few that

have been, writes: "Taken together, comparisons of teacher ratings to criterion measures other than achievement lend some criterion validity to the use of the ratings" [p. 32]. Koon (1986) depicts the implications that are relevant here:

> The key question then is the extent to which the variance shared between each criterion measure and student ratings of overall instructor effectiveness is independent of the variance shared between student ratings and the many other possible criterion measures. . . . Thus the validity of students' mean ratings of the overall effectiveness of instructors should be estimated with a multiple correlation extracted from a far more comprehensive array of learning-related measures. [P. 3rd.]

In estimating the validity of student ratings of overall instructor effectiveness, one also needs to be concerned about the validity of the achievement test used as the criterion. Ultimately, an objective test can have nearly perfect content validity only when its items are randomly selected from a strictly defined, homogenous and finite universe of items, each of which pertains to a concrete aspect of cognitive learning (Osburn, 1968). Not only is it essentially impossible to develop an objective, college-level test with unbiased contents but test validity is further reduced to the extent that higher-order aspects of cognitive learning should be included. Test scores also tend to be biased by factors such as test anxiety, test-taking ability and acculturation to the society that is implicit in a particular test. In sum, whether as a measure of learning for estimating the validity of student ratings of teaching effectiveness or for use directly in program evaluation, objective tests provide very rough estimates of the level of student learning, and are biased in several ways--features which necessarily limit their validity and restrict their correlation with measures that do not share the same patterns of error.

Further, since students are motivated to get good grades, their own efforts at study may make up for some of the inadequacies of a weak instructor (McKeachie, 1979, p. 385). Thus a poor instructor, especially in a basic course in which students must learn the fundamentals well (the type of course most often researched) might "force" students to spend more time to read or review the books assigned, to discuss topics or problem sets with other students, to increase the efficiency with which they attend to the particular subject, to make use of a help center, to go regularly to class or to discussion or lab sections, and so on. Even studies using random assignment are subject to serious consequences when the assumption of group equivalence is violated (Ross and Cronbach, 1976)--which occurs to the extent that students act to compensate for poor teaching, and when students decide to drop a course or change instructors. The upshot is that "when an objective measure of student learning is the criterion of validity, student activity to compensate for poor teaching makes student ratings of instructor effectiveness seem less valid than they actually are" (Koon, 1986, p. 4th).

When correctly re-conceptualized then, several persuasive reasons indicate that the validity of student ratings of instructor effectiveness is greatly underestimated. Indeed, when aggregated by department, student ratings of instructor effectiveness are the best measure of faculty teaching performance, which is an important criterion in academic program evaluation.

Student ratings also have been used extensively for characterizing and evaluating college/university "environments," and the information from several established instruments clearly has considerable validity (c.f., Baird, Hartnett, et al., 1980).

Pace (1985) believes that all of the available evidence indicates that college students accurately report on their activities and, with respect to their ability to rate their own learning, that college students' judgments "of what they have gained are consistent both with external evidence, when it exists, and with what we might expect in the light of their activities and interests" [p. 13]. While Pace's view may be an overstatement, there is no reason that student ratings of their experience with various aspects of academic program delivery should be more erroneous than ratings of their own learning.

Student Ratings of Academic Programs: Some Use, Much Neglect, and a Few Findings Although student ratings often have been elicited for departmental self-studies, it is still rare to find them used in a periodic, systematic, comparative, campuswide, program review process. At the University of Illinois, Champaign-Urbana, however, student ratings of their academic programs were found to be reliable and to have discriminant validity (Derry and Brandenburg, 1978; Braskamp, 1978). Student ratings also appear to have discriminated between academic programs at the college level at the University of Tennessee at Knoxville (Banta, 1985). At Northeast Missouri State University, student ratings of various aspects of their learning in general education, and in the major, are used systematically as part of a value-added approach to program evaluation and improvement (Northeast Missouri State University, 1984). Inter-institutionally, Clark, Hartnett and Baird (1976) found considerable evidence that graduate student ratings of their academic programs are valid. Yet most evaluators, and certainly most faculty members, regard student ratings of academic programs as insufficiently "expert," as purely "opinion," or as useful only for formative evaluation or needs assessment. Even authors who seem to appreciate the value of student evaluations seem to cave in to faculty or decision makers' prejudices about "credible" evidence when student ratings are challenged (e.g., Braskamp, 1982, p. 60; Centra, 1979, p. 160; Wood and Davis, 1978, pp. 26, 52). Although truth value (validity) is supposed to be the currency of the academic realm, it seems most probable that the existing distributions of prerogatives--to faculty, administrators, governing boards, and others--have played a major role in the appalling neglect even to study the potential value of student ratings of academic programs. Even the published questionnaires that currently are available (e.g., from ETS, ACT, NCHEMS) are weak in their coverage of the aspects of program delivery if not also of the aspects of student learning. If the general hypothesis here proves to be substantially correct, however, the role of student ratings will need re-thinking in centers of college and university power, as well as by adherents of many schools of evaluative thought.

Aspects of Academic Programs

Although some aspects of program delivery can be evaluated on a course-by-course basis, Dressel (1980) has argued, rightly, that programs stray, often markedly, from their professed goals, such that the individual courses

taken by a student often do not lead effectively toward the curricular results described in institutional catalogs or departmental literature. While the aspects of a curriculum constitute a large subset of the aspects of program delivery, curriculum evaluation too has been given insufficient attention (Wood and Davis, 1978; Marcus, Leone and Goldberg, 1983).

Item Contents Before proceeding to give some samples of the item wordings used in our survey(s), we will illustrate what is meant by "aspects of program delivery" and, simultaneously, provide the reader with a general idea of the contents for which undergraduate majors' ratings were elicited in this study. The aspects of program delivery include the integration of coursework and certain course sequences (e.g., necessity of prerequisites, number of required courses, effectiveness of methodology and laboratory sequences); the availability of "enough" small classes, advanced courses in the student's area(s) of interest, and sympathetic presentations of the views of different schools of thought in the field; student-faculty interaction; academic and career advising, tutoring, and orientation to the program; student workload per unit; grading rigor; the adequacy of tests; the thoroughness of feedback to students; the stimulation, challenge and quality of instruction by faculty members and by teaching assistants; whether there are felt to be enough women and ethnic minority faculty and students; program receptivity to student input; the scheduling of courses; and so on.

In addition, a few items in our surveys asked students to rate their program's contribution to their development of skills in written and oral communication, and in quantitative analysis, to their preparation for a career, and for graduate or professional school, to goals such as a broad, general education, training in critical thinking/analysis, and developing social responsibility or skills useful in the community, and to learning their subject matter, and retaining what they had learned. Table 1 below gives examples of some of the items that were used in the survey of engineering and physical sciences program students.

TABLE 1
Sample Items: Undergraduate Major Program Evaluation Questionnaire*

18. The program's system of information and advising (faculty, secretaries, catalogs, etc.) makes it easy for me to choose courses, professors, and to understand how best to meet my academic needs.
20. I have a regular faculty advisor (or otherwise know a faculty member in my program) to whom I would feel comfortable going at any time during the year.
29. Compared to work in other programs, the students in this program must do a lot of work for the number of units (credit) awarded.
34. In my opinion, there are enough women faculty in this program.
39. Some of the lower division prerequisites are not really necessary for the upper division program.
41. Advanced courses are offered in the major (or are otherwise available) in all of my areas of interest.
46. Large lecture classes can generally provide adequate learning opportunities in this field.
48. There are enough fieldwork, internship, group, lab, or other academic-

experiential course offerings in this program, and/or credit is readily available for appropriate types of volunteer experience.

50. The program's methodology, lab, or technical courses do not provide the student with enough experience in real problem-solving and analysis.

53. I will remember a lot of what I have learned in the program's courses for a long time to come.

56. I trust the faculty to take student needs and views into account in their decision-making for the program.

61.-63. My skills in written communication [oral communication, quantitative analysis] will be adequate by the time I graduate.

---

*The response options for most items are: 1= Strongly Disagree, 2= Disagree, 3= In Between, 4= Agree, 5= Strongly Agree, and 0= Don't Know or Not Applicable.

Sample Characteristics

Survey-based student ratings of 50 undergraduate major programs and of lower division programs in five colleges were gathered while the author served as Coordinator of the ASUC Academic Review Unit at U.C. Berkeley, from 1976 to 1981.* Typical items and subject matters are reported and illustrated above.

Of particular interest here are some test-retest ratings, obtained at a three-month interval in 1980, from a subsample of 56 junior and senior respondents enrolled in engineering and physical science programs. Survey responses provided by 620 students' ratings of these 12 different programs, and by 496 students enrolled in 14 social science-related programs in 1979, provide additional insight into the measurement properties of student ratings as well as a comparative context for the test-retest ratings.**

The larger samples included all of the upper division majors in smaller programs; in larger programs, every nth student, with varying nths, was chosen from a quasi-alphabetical, zip code-ordered list of majors provided by the University. Apart from the single re-mailing used with the retest subsample, double mailings were used. The response rates were fairly typical of surveys of students in large universities, ranging from 50-60% in the two larger samples and in the subsample, but varying from 37% to 72% for the 26 programs.

---

* Funding for this work was granted to the Academic Affairs Vice President of the Associated Students (ASUC) by the (former) Office of the Associate Vice Chancellor for Academic Development, upon recommendation of the Council on Educational Development (a committee of the Berkeley Division of the Academic Senate), with money provided by the State Legislature and Governor, initially granted to the University at the request of the U.C. Student Lobby (which is directed by the Student Body Presidents' Council), to evaluate and improve undergraduate education and teaching. Portions of the data analysis to be presented here were originally reported to the Berkeley campus (see Koon, 1981, especially Appendix A, pp. 15-23; and Koon, 1979).

** Footnote is on the page following.

Representativeness of the Samples The test-retest subsample consisted of every fourth respondent to the first mailing of the engineering and physical sciences survey. Time constraints were a major consideration in deciding to use only those students who had responded to the first mailing, but we also tested for differences in ratings between respondents to the first and second mailings in the larger samples (see below). In addition, among the students in the test-retest subsample, considering 88 item means (pertaining to demographic status, the academic program or student outcomes, and student goals and attitudes), there were no significant differences between the initial responses of those who did and did not (N= 49) respond subsequently to the retest.

For the larger samples, checks on the representativeness of the various groups of respondents (e.g., within programs: by gender, by ethnic identification, and by year in school; and, for the whole of the social science-related sample, by zip code/residence) revealed only a few, minor imbalances that probably had very little affect on the mean ratings.

Comparisons of the mean ratings given by respondents to the first mailing with the mean ratings by students who responded only after the second mailing also were used to estimate the extent to which non-respondents might differ from respondents. Since respondents to the second mailing would have been non-respondents without the second mailing, the assumption is that, in contrast to respondents to the first mailing, they resemble the remaining non-respondents (c.f., Kaufman and English, 1979, p. 265).

To compare the views of respondents, first mailing versus second, in the engineering and physical sciences sample, we tested the means for 84 evaluative items and 18 items pertaining to student goals and attitudes (e.g., the importance of job security as a factor in the student's choice of career or major). In five instances, the probability that the means were the same was lower than .05 (ranging from .0066 to .0429) but, with 102 tests, 5.1 instances would have been expected by chance alone. Similar results were obtained for 77 items (65 of them evaluative of the programs) in the social science-related sample. In sum, it probably is unfair to label as "significant" any of the observed differences between the item means given by first versus second mailing respondents. But even if all the non-respondents had mean ratings that were equivalent to those of the respondents to the second mailing, the largest changes in overall item

---

** [From previous page] The physical science and engineering programs include: Chemistry and Chemical Engineering (College of Chemistry), Civil Engineering, Electrical Engineering and Computer Sciences, Industrial Engineering and Operations Research, Materials Science and Engineering, Mechanical Engineering, and Nuclear Engineering (College of Engineering), and Computer Science, Geology, Mathematics, and Physics (College of Letters and Science).

The social science-related programs include: Anthropology, Development Studies, Economics, Geography, Linguistics, Political Science, Psychology, the Social Science Field Major, Social Welfare, and Sociology (College of Letters and Science), Business Administration (School of Business), and Conservation of Natural Resources, Political Economy of Industrial Societies, and Political Economy of Natural Resources (College of Natural Resources).

means, affecting perhaps five percent of the items, would be on the order of 0.15 to 0.20 scale points, or only four or five percent of the 1-5 scale length.

Methods of Analysis

Pearson product-moment correlation coefficients were calculated to provide estimates of the reliability of each of the individual raters who responded twice, indicating the extent of their agreement/disagreement (usually), on a 1-5 scale, to 58 items characterizing (primarily) the delivery processes of, and (secondarily) the contributions to learning by, academic programs. Pearson product-moment correlation coefficients also were calculated to provide estimates of the reliability of the 58 items. Pairwise deletion of missing data was used, so the actual number of responses involved in each of the calculations varies somewhat.

Second, F-tests were used to compare the variability over the three-month period in the mean ratings by the 56 students of each of the 58 items.

Third, F-tests were used to test for the reliability of differences in item means by program, within the physical science/engineering and within the social science-related samples. Significant differences between program means also would provide evidence of discriminant validity, at least in terms of the perceptual reality of (validity to) the raters. While further statistical analyses might improve the estimate of the extent to which the variability in mean ratings by program is associated with differences in program effectiveness, a well-informed institutional researcher or higher educator can eyeball the range of means to identify item results which may be influenced by the apparent character of the disciplines aggregated into the two large samples, by the "professional" versus "academic" status of some of the programs, and by the character of the individual disciplines or of their students.

In addition to the relevance to reliability and validity, these F-tests also provide a crucial context for the second step in the analyses. That is, if the item means are virtually the same from program to program in the larger samples, their stability in the test-retest subsample might be artifactual. For example, item wordings could be a problem, or quality might be at the same level from program to program.

Results

The mean Pearson product-moment correlation coefficient of test-retest (intra-)rater reliability was .655, the median .673. The mean Pearson product-moment correlation coefficient of test-retest item reliability was .569, the median .577. The distributions of both types of reliability coefficients are shown in the tables on the following page.

In order to clarify further the meaning of the item test-retest reliability coefficients, the students' behaviors also may be summarized by averaging the percentage of 1-5 responses that were observed in various cells of 58 five-by-five crosstabulations, each of which compares test versus retest responses for a particular item over the three-month period. Viewed from this perspective, 50.7% of the individual responses were exactly

TABLE 2

Distribution of Test-Retest Reliability Coefficients for 55 Engineering and Physical Science Majors Who Rated Their Academic Program on 58 Survey Items at a Three-Month Interval*

Pearson Product-Moment Correlation Coefficient Range		Frequency No.	Frequency Pct.	Descriptive Statistics	
-1.000	to -.000	0	0		
+ .000	.099	0	0	Mean Pearson R	.655
.100	.199	1	2	Standard Dev.	.142
.200	.299	0	0	Mean Pearson R Sq.	42.89%
.300	.399	2	4		
.400	.499	6	11	Median Pearson R	.673
.500	.599	9	16	Med. Pearson R Sq.	45.35%
.600	.699	15	27		
.700	.799	14	25		
.800	.899	8	14		
.900	1.000	0	0		
		55	99		

*All the items employed a 1-5 response scale. For 48 items, that scale was 1= Strongly Disagree, 2= Disagree, 3= In Between, 4= Agree, and 5= Strongly Agree. One student response was too late for inclusion in this analysis.

TABLE 3

Distribution of Test-Retest Reliability Coefficients for 58 Survey Items Evaluative of Academic Programs,* Based on Ratings by 56 Engineering and Physical Science Majors Who Responded at a Three-Month Interval

Pearson Product-Moment Correlation Coefficient Range		Frequency No.	Frequency Pct.	Descriptive Statistics	
-1.000	to -.000	0	0		
+ .000	.099	0	0	Mean Pearson R	.569
.100	.199	0	0	Standard Dev.	.134
.200	.299	3	5	Mean Pearson R Sq.	32.37%
.300	.399	4	7		
.400	.499	8	14	Median Pearson R	.577
.500	.599	19	33	Med. Pearson R Sq.	33.29%
.600	.699	12	21		
.700	.799	11	19		
.800	.899	1	2		
.900	1.000	0	0		
		58	101		

*All the items employed a 1-5 response scale. See also Table 2 and the text.

systematically, as part of the program evaluation process at every college and university. For each institution, analyses probably will be needed to estimate the range of item and rater reliabilities typically observed (although successful differentiation between programs may well tell much of that tale). If results similar to ours are observed in other studies, the validity of student ratings will hinge on the question of bias. If the measurement error in student ratings proves to be mostly random, there is little problem. It may also prove possible to control statistically or counterbalance with other measures such biases as may be observed.

Since disciplinary area is likely to be the principal source of bias or difference, simple comparisons of mean student ratings of programs within each disciplinary area/aggregation are essential in any analysis. Periodic collection of data would provide information to develop standards that are individualized for each program, and would enable the use of a time-series approach for many analyses, not only for studying prior and current student ratings of program performance, or of student learning, but for comparing such data with other measurements taken, or program changes made, over time. But in conducting any form of criterion-related analyses of the validity of the student ratings, remember that two different types of measures, though targeted at similar constructs, will seldom be closely fitted to each other, that several criterion measures and a multiple correlation often will be necessary to do justice to the meaning encompassed by rated variables,* that strict behavioral measures almost invariably have biases of their own, and that unmeasured interactions may suppress expected relationships.

Multi-institutional research on student evaluations of their academic programs (including aspects of their own learning and development) also is essential. If the power of student evaluations to differentiate between programs that was observed at Berkeley is present across institutions, just think of the value that student ratings could have, for example, on a statewide basis, in identifying the relative strengths and weaknesses of various aspects of program delivery, or of various aspects of learning. Although the validity of undergraduate student ratings of academic programs might vary to some extent by institution, when Pace and Friedlander (1982) compared undergraduate student ratings of several of their own academic and academic support-related activities with self-reported behavioral frequencies for the same activities, they found that the rated response categories "never," "occasionally," "often" and "very often" were interpreted, in terms of behavioral frequencies, in very nearly the same way by students from a diverse sample of institutions. And, except for dimensions concerned with the quality of the teaching/learning experience, Clark, Hartnett and Baird (1976) found that graduate student ratings of academic programs generally were valid when the criteria were faculty ratings and doctoral alumni ratings. The exception occurs primarily because faculty peers cannot differentiate instructional quality from the quality of faculty scholarship.

---

* Since rated variables may be fitted to almost any content, however, certain specific tests of validity (albeit with unknown generalizability) may be designed by fitting survey items closely to the (actual) content of existing measures. Clearly, it will be easier to study the validity of student ratings of their own learning than of ratings of the various aspects of program delivery.

If our hypothesis is substantially correct, as the available evidence suggests it is, the amount of relatively inexpensive but precisely-targeted information available through surveys will make student ratings of academic programs the evaluative method of choice for almost all aspects of program delivery, and possibly many aspects of student learning and development. Several measures with diverse foci, chosen from the traditionally acceptable types and with the selection varying somewhat for each periodic evaluation, would be used to complement and provide a check on the student evaluations.

If student ratings are generally valid, they can also be of immense value for program improvement, and for avoiding situations in which faculty members have identified problems but must fish for solutions because they have little idea about which aspects of program delivery are and are not effective.

And, as suggested earlier, if well-designed instruments enable students to provide reasonably valid assessments of most aspects of program delivery and of at least some aspects of their own learning, they might well also enable students to evaluate the extent to which the various aspects of program delivery, as contrasted with other sources of influence, contributed to various aspects of their learning and development. Such direct student ratings might enable the circumvention of the otherwise nearly impossible task, given current theory and strict behavioral measures, of attempting to estimate (the magnitude of) the effects of a particular cause in higher education.

Given this potential, it is appalling that some authors omit student evaluations of academic programs in describing model assessment programs--especially when they argue that "the priority should be on instructional feedback" and, at the same time, include the use of "institutional surveys" only to provide "aggregate data useful for policy decisions on matters such as perceived availability of services, financial aid, part-time work, job placement, library resources, computer facilities, and so on" (Turnbull, 1985, p. 25). This is not to say, of course, that any survey design will effectively reflect such differences between academic programs as may exist, but the student-centered item design that we used certainly seemed to do so.

References

Astin, A. W. (1980). Developing information systems about student activities and opinions. In L. L. Baird, R. T. Hartnett, & Associates, Understanding student and faculty life: Using campus surveys to improve academic decision making, pp. 68-89. San Francisco: Jossey-Bass.

Astin, A. W. (1985). Achieving educational excellence: A critical assessment of priorities and practices in higher education. San Francisco: Jossey-Bass.

Baird, L. L., Hartnett, R. T., & Associates. (1980). Understanding student and faculty life: Using campus surveys to improve academic decision making. San Francisco: Jossey-Bass.

Ball, S. (1981). Outcomes, the size of impacts, and program evaluation. In S. Ball (Ed.), Assessing and interpreting outcomes (New Directions for Program Evaluation No. 9), pp. 71-86. San Francisco: Jossey-Bass.

Banta, T. W. (1985). Use of outcomes information at the University of Tennessee, Knoxville. In P. T. Ewell (Ed.), Assessing educational

outcomes (New Directions for Institutional Research No. 47), pp. 19-32. San Francisco: Jossey-Bass.

Benton, S. E. (1982). Rating college teaching: Criterion validity studies of student evaluation-of-instruction instruments (AAHE-ERIC/ Higher Education Research Report No. 1). Washington, DC: American Association for Higher Education.

Bowen, H. R. (1974). The products of higher education. In H. R. Bowen (Ed.), Evaluating Institutions for Accountability (New Directions for Institutional Research No. 1), pp. 1-21. San Francisco: Jossey-Bass.

Braskamp, L. A. (1978, March). Student satisfaction as a useful criterion in departmental review. Paper presented at annual meeting of American Educational Research Association, Toronto, Canada.

Braskamp, L. A. (1982). Evaluation systems are more than information systems. In R. F. Wilson (Ed.), Designing academic program reviews (New Directions for Higher Education No. 37), pp. 55-66. San Francisco: Jossey-Bass.

Byrk, A. S. (1978). Evaluating program impact: A time to cast away stones, a time to gather stones together. In S. B. Anderson & C. D. Coles (Eds.), Exploring purposes and dimensions (New Directions for Program Evaluation No. 1), pp. 31-58. San Francisco: Jossey-Bass.

Centra, J. A. (1979). Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement. San Francisco: Jossey-Bass.

Clark, J. L. D. (1979). Measures of student learning. In J. A. Centra, Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement, pp. 93-118. San Francisco: Jossey-Bass.

Clark, M. J., Hartnett, R. T., & Baird, L. L. (1976). Assessing dimensions of quality in doctoral education: A technical report of a national study in three fields. Princeton, NJ: Educational Testing Service.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 51, 281-309.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Boston: Houghton Mifflin.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity and usefulness. Review of Educational Research, 41, 511-534.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., Walker, D. F., & Weiner, S. S. (1980). Toward reform of program evaluation. San Francisco: Jossey-Bass.

Derry, S., & Brandenburg, D. C. (1978). Students' ratings of academic programs: A study of structural and discriminant validity. Journal of Educational Psychology, 70, 772-778.

Dowell, D. A., & Neal, J. A. (1982). A selective review of the validity of student ratings of teachers. Journal of Higher Education, 53, 52-62.

Dressel, Paul L. (1980). Improving degree programs: A guide to curriculum development, administration, and review. San Francisco: Jossey-Bass.

Ewell, P. T. (1984). The self-regarding institution: Information for excellence. Boulder, CO: National Center for Higher Education Management Systems.

Feldman, K. A., & Newcomb, T. M. (1973/1969). The impact of college on students. San Francisco: Jossey-Bass.

Hartnett, R. T., & Centra, J. A. (1977). The effects of academic depart-

ments on student learning. Journal of Higher Education, 48, 491-507.
The Joint Committee on Standards for Educational Evaluation. (1981). Standards for evaluations of educational programs, projects, and materials. New York: McGraw-Hill.
Kaufman, R., & English, F. W. (1979). Needs assessment: Concept and application. Englewood Cliffs, NJ: Educational Technology Publications.
Koon, J. (1981). Engineering and physical science programs at Berkeley: An academic review using comparative student evaluations. University of California, Associated Students, Academic Affairs Office. (ERIC Document Reproduction Service No. ED 198 000)
Koon, J. (1979). Social science-related programs at Berkeley: An academic review using comparative student evaluations. University of California, Associated Students, Academic Affairs Office. (ERIC Document Reproduction Service No. ED 180 903)
Koon, J. (in press). Teaching evaluation: Theory and practice for a student position. Issues in Higher Education.
Kuh, G. D. (1981). Indices of quality in the undergraduate experience (AAHE-ERIC/Higher Education Research Report No. 4). Washington, DC: American Association for Higher Education.
McKeachie, W. J. (1979). Student ratings of faculty: A reprise. Academe, 65, 384-397.
Marcus, L. R., Leone, A. O., & Goldberg, E. D. (1983). The path to excellence: Quality assurance in higher education (ASHE-ERIC/Higher Education Research Report No. 1). Washington, DC: Association for the Study of Higher Education.
Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. American Educational Research Journal, 17, 219-237.
Northeast Missouri State University. (1984). In pursuit of degrees with integrity: A value added approach to undergraduate assessment. Washington, DC: American Association of State Colleges and Universities.
Osburn, H. G. (1968). Item sampling for achievement testing. Educational and Psychological Measurement, 28, 95-104.
Pace, C. R. (1985). Perspectives and problems in student outcomes research. In P. T. Ewell (Ed.), Assessing educational outcomes (New Directions for Institutional Research No. 47), pp. 7-18. San Francisco: Jossey-Bass.
Pace, C. R., & Friedlander, J. (1982). The meaning of response categories: How often is "Occasionally," "Often," and "Very Often"? Research in Higher Education, 17, 267-281.
Ross, L. & Cronbach, L. J. (Eds. [for A Task Force of the Stanford Evaluation Consortium]). (1976). Book review [of Handbook of evaluation research]. Proceedings of The National Academy of Education, 3, 81-107.
Solmon, L. C. (1981). A multidimensional approach to quality. In T. M. Stauffer (Ed.), Quality--higher education's principal challenge, pp. 6-14. Washington, DC: American Council on Education.
Turnbull, W. W. (1985). Are they learning anything in college. Change, 17(6), pp. 22-26.
Wood, L. & Davis, B. G. (1978). Designing and evaluating higher education curricula (AAHE-ERIC/Higher Education Research Report No. 8). Washington, DC: American Association for Higher Education.